PR #24772 完整报告

sgl-project/sglang

Fix PD bootstrap failure handling

合并时间：2026-05-10 19:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24772

执行摘要

- 一句话：修复 PD 引导失败时状态异常与属性错误
- 推荐动作：建议合入。该 PR 是典型的边界条件修复，改动量小且逻辑正确，经 review 确认可读。对于关注 PD 部署稳定性的团队，值得了解该修复以避免生产环境中的无声错误。

功能与动机

在 PD（Prefill/Decode）分离部署场景下，若预填充服务器 bootstrap 失败（如 prefilling 端宕机），_setup_bootstrap_infos 虽将 conclude_state 标记为 Failed，但未立即返回，导致下游 update_status(WaitingForInput) 覆盖失败状态，且 bootstrap_infos 可能因未定义而触发 AttributeError。PR body 明确说明动机：修复 bootstrap info 获取失败时的异常处理。

实现拆解

1. 在 _setup_bootstrap_infos 中失败路径显式置 None：当 _get_bootstrap_info_from_server 返回 None 且 conclude_state 设为 Failed 后，在 return 前添加 self.bootstrap_infos = None，使下游检查时触发 None 检查而非 AttributeError。
2. 在 _setup_bootstrap_infos 调用后立即检查失败状态：在 kv_mgr.update_status(self.bootstrap_room, KVPoll.WaitingForInput) 前插入 if self.conclude_state == KVPoll.Failed: return，避免失败请求被错误地转为 WaitingForInput。
3. 修复涉及的文件：仅 python/sglang/srt/disaggregation/common/conn.py，改动量 +3/-0。
4. 回退多余变更：提交历史显示作者先尝试引入 data_parallel_controller.py 中的外部 DP rank 路由跳过逻辑，经 review 讨论后认为由 #23882 处理更恰当，随即 revert 该部分，最终 PR 仅保留 conn.py 的 3 行核心修复。

关键文件：

- python/sglang/srt/disaggregation/common/conn.py（模块 连接模块；类别 source；类型 core-logic；符号 PrefillStatePoll.init, PrefillStatePoll._setup_bootstrap_infos）：唯一的修改文件，包含 PD bootstrap 失败的恢复逻辑：置空 bootstrap_infos 并在失败后跳过 WaitingForInput 状态更新。

关键符号：PrefillStatePoll.init, PrefillStatePoll._setup_bootstrap_infos

关键源码片段

python/sglang/srt/disaggregation/common/conn.py

唯一的修改文件，包含 PD bootstrap 失败的恢复逻辑：置空 bootstrap_infos 并在失败后跳过 WaitingForInput 状态更新。

```
# python/sglang/srt/disaggregation/common/conn.py
# 在 PrefillStatePoll.init 方法中，
# 调用 _setup_bootstrap_infos 之后检查 bootstrap 是否失败
self.prefill_dp_rank = prefill_dp_rank
self._setup_bootstrap_infos()
# 新增：若 bootstrap 已标记失败，则跳过 WaitingForInput 状态更新
if self.conclude_state == KVPoll.Failed:
    return
self.kv_mgr.update_status(self.bootstrap_room, KVPoll.WaitingForInput)

def _setup_bootstrap_infos(self):
    """Fetch bootstrap info from prefill server for all ranks."""
    all_bootstrap_infos = []
    for target_cp_rank in self.target_cp_ranks:
        bootstrap_key = f"{self.bootstrap_addr}_{self.prefill_dp_rank}_{target_cp_rank}_{self.target_
        tp_rank}"
        if bootstrap_key not in self.kv_mgr.connection_pool:
            bootstrap_infos = []
            for target_tp_rank in self.target_tp_ranks:
                for target_pp_rank in reversed(self.target_pp_ranks):
                    bootstrap_info = self._get_bootstrap_info_from_server(...)
                    if bootstrap_info is not None:
                        bootstrap_infos.append(bootstrap_info)
                    else:
                        self.kv_mgr.record_failure(self.bootstrap_room, ...)
                        self.conclude_state = KVPoll.Failed
                        self.kv_mgr.update_status(self.bootstrap_room, KVPoll.Failed)
                        self.bootstrap_infos = None # 新增：避免下游 AttributeError
                        return
            # ... 成功路径
        # ...
```

评论区精华

Reviewer ShangmingCai 对提交 702e0c5 中的 data_parallel_controller.py 变更提出质疑，认为在 PD-prefill 模式下直接跳过外部 routed_dp_rank 会破坏 bootstrap_room 与 DP rank 的映射契约。作者 yhyang201 同意该观点，并指出 main 分支已通过 #23882 更好地处理该问题，因此回退了该 commit。最终 reviewer 确认“common backend modification looks good”并批准。

- 外部 DP rank 路由跳过逻辑是否应合入 (design): 回退 data_parallel_controller.py 的变更，PR 仅保留 conn.py 的核心修复。

风险与影响

- 风险：风险极低。改动仅 3 行，逻辑清晰：失败路径提前返回并置空引用。可能存在的风险是：其他依赖 _setup_bootstrap_infos 中失败后仍执行后续逻辑的地方未发现，但代码显式 return 已避免执行路径扩散；此外 self.bootstrap_infos = None 可能被下游代码 assert len(self.bootstrap_infos) > 0 直接捕获，但失败路径已提前返回，不会到达断言。
- 影响：直接影响 PD 分离部署模式下 bootstrap 失败的请求：此前会导致 AttributeError 或状态错误标记，修复后将正确将请求标记为 Failed。影响范围仅限于 conn.py 中的 PrefillStatePoll 初始化流程，不涉及其他模块。由于未涉及配置变更或 API 调整，对用户透明。
- 风险标记：核心路径变更

关联脉络

- PR #23882 Handle external DP rank routing in PD mode: 该 PR 被作者引用为更好处理外部 DP rank 路由的方案，也是回退 data_parallel_controller.py 变更的依据。
- PR #24878 [Bug] Add dsv4 state_type branch to mooncake disaggregation: 同为 PD/disaggregation 修复的近期 PR，属同一模块，表明团队正持续加固 PD 稳定性。
- PR #24861 [Utils] Refactor device cache emptying: 同一作者 yhyang201 的近期 PR，表明其在 PD 及设备管理方面的持续贡献。